Finding Gaps in Data to Guide Development of a Radiation Threat Adjudication System

Nicholas Gisolfi¹, Madalina Fiterau¹, Artur Dubrawski¹, Saswati Ray¹, Simon Labov², Karl Nelson²
¹Auton Lab, Carnegie Mellon University, Pittsburgh, PA ²Lawrence Livermore National Laboratory, Livermore, CA

Motivation
Modeling nuclear threats in synthetic data challenges human data engineers to include all relevant niches of the feature space. High-dimensional synthetic data is prone to omissions of meaningful information which may be used to improve a trained model's accuracy at a classification task.

We aim to provide a framework which presents insufficiencies of training data in a user-friendly manner, allowing data engineers to inject data needed to fill gaps in the feature space.

Data
• Multiple Classes
• 113 Features
• Over 50K Samples

Classes include:
1. Non-Emitting sources
2. Emitting sources posing a threat
3. Emitting sources explainable by naturally occurring radioactive materials

The Iterative Build Process

The Learning System – Random Forest

Build random forest using k-fold cross validation which admit diagnostics

Diagnostic 1 – Agreement Score
• Describes the extent to which predictions made by all trees agree.
• Optimal, all trees in the forest reaching the same classification label for a given sample.

Diagnostic 2 – Inbounds Score
• Quantifies whether or not a query falls within a range of values that has been observed by a tree in the random forest during training.
• Optimally, all trees have seen a sample of similar feature values during training.

Regression Based Informative Projection Recovery

Search subspaces to find projections where data is most separable

Overview of Algorithm
for each 2D subspace in the feature space
Train classification model
for sample in training set
Evaluate loss function
Associate each point with ideal projection
Visualize most populated projections

Non-Parametric Loss Estimator
• Ratio of distances between a query sample and samples of similar and different classes
• Identifies irregular gaps

Parametric Loss Estimator
• Distance to a decision boundary
• Identifies structured gaps

Conclusions
• Framework generates visualizations which allow engineers to make changes necessary to improve synthetic data generation.
• By resolving gaps in training data, model classification performance may improve.
• Nonparametric loss function finds irregular gaps.
• Parametric loss approach reveals structured gaps and divides in the data, allowing users to easily identify adjustments to data generation that will improve model accuracy.

References